SUMMARY OF RESEARCH

DEUTEROSTOME EVOLUTION: LARGE DATA SET ANALYSIS

NAG2-1399

PI DANIEL JANIES

AMERICAN MUSEUM OF NATURAL HISTORY
79TH STREET AT CENTRAL PARK WEST
NEW YORK, NEW YORK 10024

April 1, 2000 - January 31, 2004

Deuterostome Evolution: Large Data Set Analysis
NAG2-1399
Submitted by Daniel Janies and Ward Wheeler

It is our pleasure to communicate to NASA our final report on NAG2-1399. This grant, entitled, "Deuterostome evolution: Large Data Set Analysis" begun 04/01/2000 and ended 04/01/2003. Supplemental funding was provided through 01/31/04.

The PI is Daniel Janies <janies-1@medctr.osu.edu> who was a research scientist working with Ward Wheeler <wheeler@amnh.org> a Curator in the Division of Invertebrate Zoology of the American Museum of Natural History, Central Park West at 79th Street, New York, NY 10024 for most of the grant period. As of January 16, 2003, Daniel Janies has been an assistant professor in the Department of Biomedical informatics of the Ohio State University, 333 W. 10th Ave. Columbus, OH.

This award allowed us to develop novel hardware for phylogenetics, collect genomic data and produce several phylogenies of deuterostome organisms, communicate the results publicly, release software into the public domain, publish textbooks and papers, and prepare for the next research projects. There are no resulting subject inventions to report. We review these activities in three sections: 1) Hardware and software and development, 2) Evolutionary biology research and 3) Our proposed future direction, predictive analysis of pathogens in support of the NASA mission.

1) Hardware and Software Development:

In this opening period of this award we focused on synergistic development of hardware and software to produce evolutionary trees from large data sets. Using off-the-shelf PC components we have built an internationally ranked 800-processor computer for a fraction of the cost of a traditional supercomputer. The AMNH cluster has become central to the work of more than 30 researchers in paleontology, zoology, and astrophysics. Many use our in-house software POY, an integer-intensive DNA sequence alignment and direct optimization software package.

The central problem in evolutionary biology consists of reconstructing the ancestry of organisms and changes in their features (genomes, anatomy) through time. Biologists test evolutionary hypotheses by comparing phylogenetic trees that depict organismal relationships.

Mathematically speaking, phylogenetic trees are connected acyclical graphs. The leaves of these graphs represent biological species and their features and the internal nodes are interpreted as hypothetical ancestors containing combinations of features that link descendants while minimizing the number of evolutionary changes assumed. The most frequently used methods in phylogeny aim to optimize an objective function (tree length, likelihood, or posterior probability) on the space of possible trees. Regardless of the objective function chosen, the number of possible trees increases combinatorially with the number of organisms being analyzed. Moreover, today's biologically interesting datasets are typically comprised of tens to hundreds of organisms and kilobases to megabases of DNA. Thus the number of possible trees is prohibitively large for an exhaustive evaluation of evolutionary scenarios. Thus for biologically

interesting datasets, the problem of phylogenetic tree search is compute bound and must be approached through heuristics and parallelism.

Our core algorithm for comparative genomics, tree-based DNA alignment, is an effective means of simultaneously evaluating many possible scenarios of DNA sequence homology and evolutionary relationships among organisms. We have successfully combined parallel tree-search and tree-alignment heuristics with self-built computing clusters comprised of inexpensive commodity PC components.

Most phylogenetic studies are based on standard strategy (randomization and hill climbing). This strategy combined with parallelism only allows an investigator to examine more random additions of taxa or ratchet replicates concurrently. When considering the importance of hundreds to thousands of replicates in current phylogenetic theory it is impressive that the overall search time for hundreds of taxa is reduced from years to weeks. However we developed genetical algorithms that are reducing search times for large datasets dramatically, from weeks to days. In 2001 we were ranked 9th in the world for computing clusters and were the most cost efficient. As a result of the efficiencies we demonstrated from synergistic hardware and software development, cluster computing has become widely adopted among biologists working on a wide variety of problems. Phylogenetic analyses are becoming critical tools for numerous disciplines in biomedicine. These fields include genomics, ecology, physiology, pharmacology, epidemiology, developmental biology, and forensics.

Press coverage related to hardware and software development:

2001 "The Do-It-Yourself Supercomputer", *Scientific American* (New York), August. http://www.sciam.com/2001/0801issue/0801hargrove.html

2001 "Supercomputing from Scratch", *Genome Technology* (New York), July. http://research.amnh.org/users/djanies/genometech.jpg

2001 "Growing PC's into Supercomputers, all in a row", *The New York Times*, June 21. http://www.nytimes.com/2001/06/21/technology/21NEXT.html

2001 "Finding Power in Numbers", *The Record* (New Jersey). June 15. http://research.amnh.org/users/djanies/NJrecord.jpg

2001 "Top Computing Clusters", http://clusters.top500.org (Germany). http://clusters.top500.org/db/entry.php3?id=168

2000 "Homegrown Computer Roots out Phylogenetic Networks", *Nature* (London), March 16 http://research.amnh.org/users/djanies/Nature_404.214.pdf

Publications related to hardware and software development:

In press Wheeler, Janies et al., Phylogenetic Systematics using POY. Harvard University Press.

2004 Wheeler, W., Janies, D., DeLaet, J., Parallel computing and sequence alignment. McGraw-Hill 2004 Yearbook of Science & Technology.

2003 Janies, D. and W. Wheeler. poy3.pdf. A primer for evolutionary analysis of combined DNA and anatomy data in cluster computers. http://research.amnh.org/users/djanies/poy3.pdf

2002 Janies, D. and W. Wheeler. Theory and practice of parallel direct optimization, *in* Techniques in Molecular Systematics and Evolution. R. Desalle, G. Giribet, and W. Wheeler (eds). Birkhauser Verlag, Basel. pp. 115-23.

2001 Janies, D and W. Wheeler. Efficiency of parallel direct optimization. *Cladistics.* 17: S71-S82. http://www.idealibrary.com/links/doi/10.1006/clad.2000.0160/pdf

2) Evolutionary biology:

Although we have enabled work across biological disciplines via hardware and software development in phylogenetics, our specific evolutionary work concerns the Deuterostomes. Within the history of the Deuterostomes significant morphological innovations in the nervous system (neural tube) and in body structure (notochord and lateral muscle bands) have resulted in one of the most important animal radiations in the history of Earth. The combination of complex, intelligent behavior with dexterity and mobility has allowed vertebrates to sit at the top of most ecosystems, to develop complex communication and social organization, and to explore beyond their planet of origin. Our work has produced several phylogenies of the basal member of the deuterostomes: the hemichordates and echinoderms, as well among chordates.

Press coverage related to evolutionary biology:

2002 "All in the Family: Putting every creature with its kin on the tree of life" *U.S. News and World Report* (Washington) June 3. http://www.usnews.com/usnews/issue/020603/misc/3tree.htm

Recent publications related to evolutionary biology:

In press Kerr, A.M., D. Janies, R. M. Clouse, J. Kuszak and J. Kim. Molecular phylogeny of coral-reef sea cucumbers (Holothuroidea: Aspidochirotida) based on 16S mt rDNA sequence. *Marine Biotechnology.*

Submitted Clouse, R., D. Janies, and A. M. Kerr. Resurrection of Bohadschia bivittata from B. marmorata (Holothuroidea: Holothuriidae) Based on Behavioral, Morphological, and Mitochondrial DNA Evidence. *Zoology*

2003 Janies, D. Reversibility in life cycle and larval evolution among echinoderms. *Cladistics* 19:154.

2001 Frost, D., R. Etheridge, D. Janies, and T. Titus. Total evidence, sequence alignment, evolution of polychrotid lizards, and a reclassification of the Iguania (Squamata: Iguanania). *American Museum of Natural History, Novitates.* 3343. 38 pp.

2001 Giribet, G., D. Janies, and W. Wheeler. Introduction. *Cladistics.* 17: S1-S2. http://www.idealibrary.com/links/doi/10.1006/clad.2000.0153/pdf

2001 Janies, D. Phylogenetic relationships of extant echinoderm classes. *Canadian Journal of Zoology.* 79:1232-1250. http://research.amnh.org/users/djanies/janies.pdf

2001 Frost, D., D. Janies, P. le Fras Mouton, T. Titus. A molecular perspective on the phylogeny of the girdled lizards (Cordylidae, Squamata). *American Museum of Natural History, Novitates.* 3310. http://nimidi.amnh.org/pubs/novitat5.html

2000 Blake, D., D. Janies, and R. Mooi. Evolution of Starfishes: Morphology, Molecules, Development, and Paleobiology. *American Zoologist.* 40: 311-315.

3) Predictive Analysis of Pathogens in Support of the NASA Mission:

History repeats itself. Phylogenetics is the retrospective analysis of biological change and adaptation over time. Originally, this field was considered relevant only to taxonomic and evolutionary studies. However, because phylogenetics is able to detect not only changes that discriminate among lineages but also changes that have occurred many times in parallel, phylogenetic trees provide significant predictive power.

Specifically, information on the direction, frequency, order, and reversibility of genomic changes that correlate with pathogenic phenotypes are vital to understand the emergence of pathogenicity and how microbial life will adapt in space. As long-term human exploration missions are planned, a better understanding of the emergence of pathogenicity among viruses and bacteria associated with space and earth travel is vital for NASA crews, the public, the NASA workforce, and high-value equipment and property. This need has prompted us to extend our software to genomic level analysis. Genomic rearrangement, within and between microbial lineages, is a biologically important but analytically under-explored level of biological complexity. We have a working set of novel algorithms to simultaneously search for the most efficient series of substitutions and insertion-deletions at the level of the nucleotide and origins, losses, translocations, and inversions of loci within genomes. The output is more than just a phylogeny but also a hierarchical database of shared genomic changes between ancestral and descendent

chromosomes or plasmids. Conversely, our work is producing unique information vital to design sensitive PCR, microarray, and immunological diagnostic reagents and devices.

Press coverage related to predictive analysis of pathogens:

2003 "SARS may be mammal-bird merger", Nature Science Update, 19 December 2003, http://www.nature.com/nsu/031215/031215-9.html

2003 "SARS gene built from 2 viruses", Toledo Blade (Ohio), October 4.

Recent publications related to predictive analysis of pathogens:

2004 Janies. D. Evolution of SARS associated coronavirus. *Cladistics*. 20:86.

2003 Wheeler, W. Chromosomal Character optimization and arthropod phylogeny. *Cladistics*: 19:161.